Technical Assignment 4:

Devin W. de Silva
Faculty of Information
School of Graduate Studies, University of Toronto

INF2178: Experimental Design for Data Science
Professor Shion Guha
April 5, 2024

The dataset we are working with examines cognitive health changes across a sample population of demented and nondemented patients identified via MRI scans over time. It includes demographic variables (age, sex, education), cognitive performance indicators, as measured by the MMSE score, CDR score and a range of other neurological data such as eTIV, nWBV, and ASF. For my analysis, I want to focus more on the how dataset provides us insights into the patterns of progression of dementia, and whether educational factors play a role in affecting cognitive performance changes. My mixed-effects ANOVA analyses would focus on the following research questions[1]:

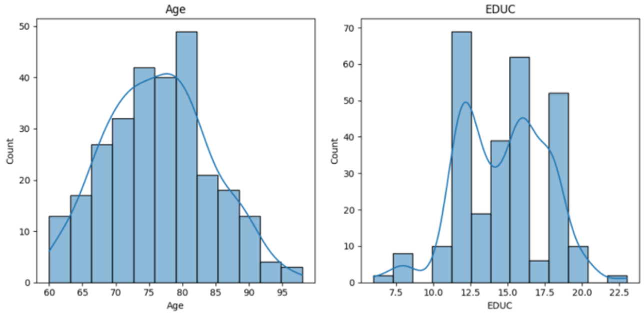
- **Do patients with dementia show a different pattern of cognitive decline over time compared to patients without dementia?** We use 'Group' as the fixed effect to compare cognitive decline (MMSE score) between groups, with 'Subject ID' as a random effect to account for repeated measures.

- **Does the level of education affect the rate of cognitive decline in patients with dementia, taking into account individual differences?** We use 'EDUC' as a fixed effect to see if education level impacts the rate of MMSE score changes in the dementia group, with 'Subject ID' as the random effect to account for within-subject correlation.

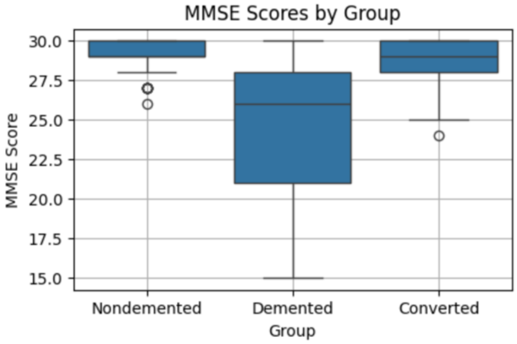
After loading the dataset, I imported all the necessary libraries and gathered some general information about the data. We have a total of 16 columns and 293 rows. Among them, SES has 15 null values and MMSE has one single null value. After dropping them, I confirmed that there are no more null values left in the dataset. Then, I proceeded to learn about the mean and

[1] I included a third question and model in the code but given the page limit, I wasn't able to discuss it at length. The third question was: Among patients without dementia, is there a relationship between socioeconomic status and cognitive performance over time, taking into account individual differences?

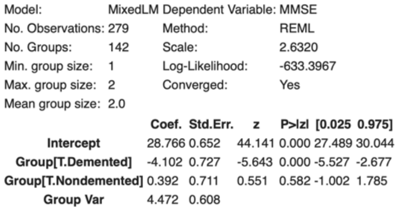
spread of the dataset using df.describe(). I selected the two most pertinent data (Age and EDUC) for my analysis and created histograms to identify their spread. As we can see, none of these data seem to be exactly normally distributed (Although Age seems markedly less skewed than the other two).



The chart below identifies how the demented group shows a much lower median score and a much larger spread of scores.



With this knowledge, I began by exploring **the first Mixed Effects ANOVA: Cognitive decline (MMSE Scores) over time, comparing groups with and without dementia.** The null hypothesis is that there is no significant difference in scores between demented and nondemented groups.

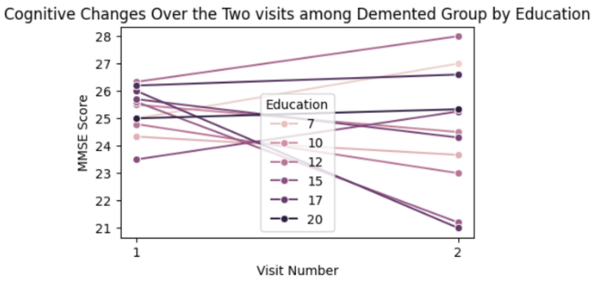


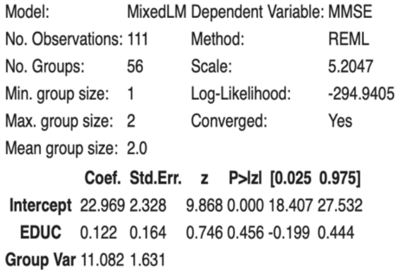
```
Model:              MixedLM Dependent Variable: MMSE
No. Observations: 279       Method:             REML
No. Groups:         142     Scale:              2.6320
Min. group size:    1       Log-Likelihood:     -633.3967
Max. group size:    2       Converged:          Yes
Mean group size:    2.0
                        Coef.  Std.Err.   z     P>|z| [0.025  0.975]
        Intercept       28.766 0.652    44.141  0.000 27.489  30.044
   Group[T.Demented]    -4.102 0.727    -5.643  0.000 -5.527  -2.677
Group[T.Nondemented]    0.392  0.711    0.551   0.582 -1.002  1.785
        Group Var       4.472  0.608
```

There are a couple of things to note here. First of all, the intercept is at 28.766, which is the estimated average MMSE score for the nondemented group. Group[T.Demented] has a negative score of -4.102, meaning that the demented group scores are on average 4.102 points lower on the MMSE than the 'Nondemented' group. Crucially, the p-value for the 'Demented' group is smaller than the significance level of 0.05, suggesting that the difference in MMSE scores between the demented and nondemented group is statistically significant. Several other values confirm this. The Z-value of -5.643 shows a strong effect in the negative direction, and a 95% confidence interval (-5.527, -2.677) remains below zero. Lastly, the group variance is also substantial at 4.472, showing that there remains evident variability in the MMSE scores that is attributed to differences between subjects in the same group.

Mixed ANOVA Question 2: Impact of education on rate of cognitive decline (in terms of MMSE Scores) in patients with dementia. The null hypothesis is that the level of education does not have a significant impact on the rate of cognitive decline.

I first created a chart showing the cognitive changes over the two visits for demented patients. The spread of the scores seems much larger by visit 2, and some patients didn't even experience a decline at all. Also shown here are the mixed effects ANOVA results:

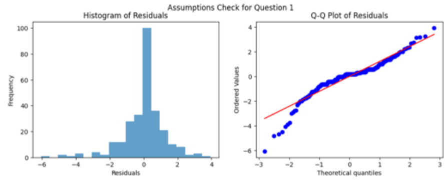


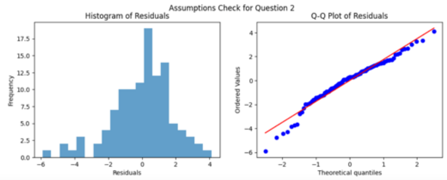


```
Model:             MixedLM Dependent Variable: MMSE
No. Observations:  111     Method:             REML
No. Groups:        56      Scale:              5.2047
Min. group size:   1       Log-Likelihood:     -294.9405
Max. group size:   2       Converged:          Yes
Mean group size:   2.0
            Coef.  Std.Err.   z    P>|z| [0.025 0.975]
Intercept  22.969  2.328    9.868  0.000 18.407 27.532
  EDUC      0.122  0.164    0.746  0.456 -0.199  0.444
Group Var  11.082  1.631
```

The intercept of 22.969 represents a theoretical average MMSE score for a hypothetical individual with an education level of zero and serves as the baseline. The EDUC coefficient of 0.122 does indeed suggest that for each additional year of education, there's an average increase of 0.122 in the MMSE score. However, this effect is not statistically significant. This is because the p-value of EDUC is at 0.456 which is much larger than 0.05. Finally, with a Group Variance of 11.082, we can tell that there is substantial variability in MMSE scores within the demented group that cannot be explained by education levels alone.

The next step is to validate the assumptions of the two mixed effects ANOVA tests we conducted. First, I plotted the above Histogram and QQ plot for Residuals of the two tests





As shown, neither of the two tests shows an exact bell-curved histogram, nor do the quantiles of their residuals completely fit on the linear regression line, especially near the tails. The second test nonetheless seems to be slightly more aligned than the first. I also defined a function to run the Shapiro-Wilks test to confirm the results. For the first model, with a test statistic of 0.91 (below 1), and a p-value much smaller than 0.05, residuals for the first test do not fit a normal distribution. For the second, the code gives a statistic of 0.9647, which is closer to 1, but the p-value is still below 0.05, meaning that it deviated from the normality assumption marginally.

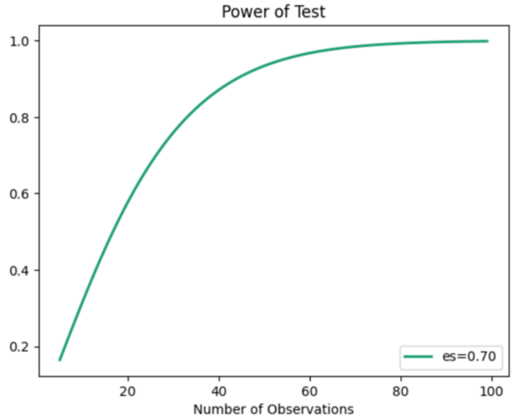
I also ran relevant Levene's tests to understand whether the homogeneity of variances assumption stands. For the first test, it provided a large statistic of 36.2074, with a very small p-value of < 0.001, suggesting that the assumption is violated. For the second model, its Levene's test statistic is much lower at 1.3627, and the p-value is 0.21 which is larger than 0.05, meaning that there is no strong evidence against equal variances across groups, thereby the assumption stands.

Conclusion

In short, the results from our analysis largely confirmed the statistically significant difference in cognitive decline between demented and nondemented groups. The demented patients do show evidently greater decline, even when individual differences are considered. It is clear that we need to targeted support and interventions for people with dementia because they indeed generally suffer from a more rapid rate of cognitive decline. The second test shows that education level itself does not significantly affect the rate of cognitive decline among demented patients. Thus, we shouldn't emphasize educational attainment too much when helping patients, and instead focus on other factors like socioeconomic conditions on cognitive decline (I have run a third test on this subject in the code). We also need to note that the normality of residuals assumption is violated in both models, and the homogeneity of variance assumption is violated in the first, which means that we need to be cautious about the potential biases within the dataset. With enough time, we can explore further into other elements of the dataset. By doing that, we would be able to gain even deeper insights into the factors causing dementia's progression and how we can mitigate them.

Sample Size Calculation and Visualization

The final part of the code relates to calculating a t-test sample size by inputting an effect size of 0.7, an alpha of 0.05, and power of 0.91. As the code shows, this figure is around 45.45. To confirm this, we created a power analysis plot, shown below. As we can see, the chart shows a logarithmic, positive correlation between the number of observations and the power of the test. If we trace 0.91 on the y-axis, our number of observations roughly lines up to 45.45 observations. Since sample sizes must be whole numbers, the lowest number of individuals we need to achieve the power we need is 46.


Power of Test
Number of Observations
es=0.70